DOI：10.13998/j.cnki.issn1002-1248.23-0300

基于 HARP 框架的农业知识图谱表示模型研究

陈彩铭[1]，冯建中[1*]，白林燕[2,3]，王　剑[1]，谢能付[1]，邹　军[1]

（1. 中国农业科学院农业信息研究所，北京 100081；2. 中国科学院空天信息创新研究院 数字地球重点实验室，北京 100094；3. 可持续发展大数据国际研究中心，北京 100094）

摘　要：[目的 / 意义]随着农业知识图谱数据规模的增长，图谱的节点和关系复杂度不断提升，这对其训练和表示提出了新的挑战。在此背景下，探索如何在保全知识图谱结构的同时降低资源消耗并加快嵌入速度具有重要的研究和应用意义。[方法/ 过程]针对这一问题，本研究提出了一种基于 HARP 框架的农业知识图谱层次表示模型。该模型利用农业知识图谱的层次性特征，采用一种改进的基于关系路径随机行走策略，有效地保留了图谱中节点的层次性和非对称关系结构。[结果 / 结论] 1)与 HARP 框架相比，使用 LEIDEN 的 HRWP 模型能更好地保留空间结构，并快速收敛了速度；2)采用 HRWP 的融合模型训练时间基本小于二者训练时间总和，且对原算法时间复杂度影响较小；3)结合 HRWP 的传统算法各指标平均提高 2%，非神经网络模型有显著提升。综上，认为模型可以准确表示农业知识图谱并有效缩短训练时间。

关键词：知识图谱；随机游走；表示学习；HARP 框架

中图分类号：TP391.1；S126　　文献标识码：A　　文章编号：1002-1248（2023）08-0066-12

引用本文：陈彩铭，冯建中，白林燕，等. 基于 HARP 框架的农业知识图谱表示模型研究[J]. 农业图书情报学报，2023，35(8)：66-77.

1 引　言

知识图谱是一种有效反映现实信息的知识结构，已被广泛应用于知识工程各领域[1,2]，通常用三元组(h,r,t)的形式表示，其中 h、t 分别表示头和尾实体，r 表示二者的关系。鉴其具备很好的信息组织和推理能力，知识图谱现已成为知识领域智能服务的重要基础设施之一[3]。然而，这种基于三元组的存储结构通常会受到数据稀疏的影响，在语义计算或关系推理时效果并不理想[4]。因此，在实际使用中通常利用知识表示学习将实体和关系表征为低维稠密向量，进而提升知识获取、

收稿日期：2023-05-16

基金项目：国家科技创新 2030 新一代人工智能重大项目课题“农业智能知识服务平台”（2021ZD0113702-02）；新疆生产建设兵团（重点领域）科技攻关计划项目“昆玉市‘互联网 +’的智慧农业集成示范应用技术研究”（2019AB002）；中国农业科学院科技创新工程项目（CAAS-ASTIP-2023-AII）

作者简介：陈彩铭（1998- ），男，硕士，研究方向为农业知识图谱及应用。白林燕（1981- ），博士，研究方向为地理遥感技术等。王剑（1976- ），博士，副研究员，研究方向为农业专业信息搜索理论与技术等。谢能付（1975- ），博士，研究员，研究方向为区块链农业应用、大规模农业知识处理，农业智能计算等。邹军（1997- ）男，硕士，研究方向为农业数字孪生

* 通信作者：冯建中（1971- ），男，博士，研究员，研究方向为信息技术与数字农业等。E-mail：fengjianzhong@caas.cn

融合和推理的性能。

随着数据组织技术不断发展和知识图谱规模的进一步增长，现阶段对大规模图谱训练产生了更高要求。传统知识图谱表示算法不仅需要大量的训练时间，而且还会消耗海量内存资源[5,6]。另外，现有大型图谱快速学习模型往往过于关注局部而忽视了长距离的全局信息，进而导致嵌入结果无法揭示语义层级等重要关系[7]。为了对大规模知识图谱的表示学习模型进行快速训练，同时保留原图中概念的层级结构，实现更高效的图谱表示学习，本文提出基于网络层次表示学习框架（Hierarchical Representation Learning for Networks，HARP）的知识图谱快速分层游走学习模型（Hierarchical Random Walk Representation Learning Model，HRWP）。模型使用分层随机游走实现知识图谱的初始表示。同时，考虑到知识图谱中关系学习的问题，模型将关系嵌入视为头实体与尾实体的分别嵌入。在实现上，模型采用阻塞随机游走（Frustrated Random Walk，FRW）实现采样，有效避免了节点沉没并实现关系的不对称嵌入，使模型能更好地与其他学习模型融合。

综上所述，本文的贡献主要体现在以下 3 点：①提出了一种大型农业图谱快速学习的模型，拓宽了基于随机游走图嵌入算法的应用场景；②通过分层游走与改进采样更好地捕捉实体间的层级关系和非对称语义，优化了现有模型关系学习不足的问题；③在真实数据集上进行定量实验，验证了本文提出的农业知识图谱表示学习快速训练模型的性能。实验结果表明，该模型可以有效地集成到包括 TransE、RotatE 等表示学习算法中，为不同应用场景中的下游任务提供更好的表示结果。

2　背景介绍

节点位置表示可以反映图的结构信息，实现标签传播[8]、项目推荐[9]和主题搜索等功能，是图研究的基本问题。现阶段网络图包括海量边和节点，因此直接操作会耗费过多资源。目前的解决方法是将高维稀疏的网络投影到低维稠密空间，即图嵌入（Graph Embedding），如图 1 所示。知识图谱的嵌入也被称为知识表示学习（KRL）。相较于 one-hot 编码，知识表示学习能显著提升图谱上作业效率。知识图谱嵌入通常包括 3 个步骤：①定义实体与关系的形式空间。实体一般被定义为连续向量，考虑到不确定性也可采用多元高斯分布建模[5]；关系可以表示为向量、矩阵、张量以及高斯分布等。②为三元组(h, r, t)定义评分函数，一般定义为距离或语义匹配度。③最大化置信度训练学习实体及关系向量的表示。

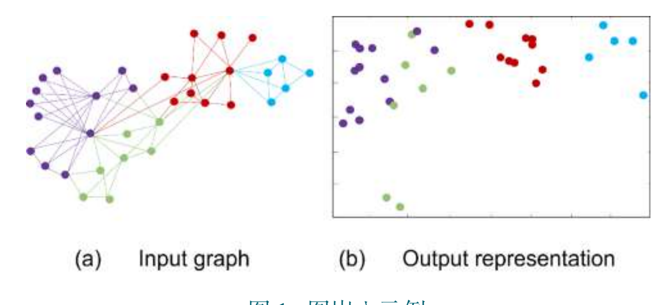


图 1　图嵌入示例

Fig.1　Diagram of graph embedding

随着数据科学的发展和大数据、云计算等技术的应用，学术及工业界针对不同应用场景提出不同的知识图谱嵌入模型，一般可分为基于翻译模型、基于语义匹配和基于神经网络的模型。翻译模型将实体 h、t 表征为向量，关系 r 表示为连接 h、t 的规则：TransE 将实体和关系表示为同一空间中的向量，关系被解释为头尾实体间的平移向量[10]；TransF 和 TransA 等模型放宽了 $h+r\approx t$ 的基础假设，使得嵌入结果更加灵活[11,12]；RotatE 将实体和关系建模到复数空间，将关系描述为复数域空间的旋转变换，对非对称关系和关系组合的表示上具备良好的效果[13]。语义匹配模型采用基于相似性的打分函数，通过匹配实体和关系在向量空间的潜在语义衡量事实成立的可能性：RESCAL 模型用向量表示实体，用矩阵表示关系，通过自定义的打分函数捕捉三元组内部的交互关系[14]；DistMult 通过限制关系矩阵为对角矩阵对 RESCAL 模型简化[15]；ComplEx 在 DistMult 的基础上引入复数空间并利用非对称打分函数更好地建模非对称关系[16]。此外，许多知识图谱嵌入也引入了神经网络的方法：R-GCN 首次将 GCN 引入到图谱的关系表示学习中[17]；SACN 引入了基于

DOI：10.13998/j.cnki.issn1002-1248.23-0300

ConvE 的加权卷积网络 WGCN[18]。

在实际应用中，NELL（Never-ending Language Learning）项目将基于随机游走的路径排名算法（PRA）作为关系推理模块[19]；Google 的 Knowledge Vault 项目利用潜在因素模型和随机游走模型结合的混合方案实现知识评估任务[20]。可见，随机游走模型具备并行运算的条件，更适合大规模知识图谱。因此，本文采用随机游走的思路可以满足实际应用中对大规模知识图谱的快速训练要求。

3 可行性分析

分层嵌入能有效保留知识图谱层次信息，对理解实体关系和知识图谱补全等任务具有重要意义[7]。分层嵌入使用低维向量降低计算复杂度以提高大规模知识图谱处理效率。同时，将实体和关系映射到低维空间的特性增加了表示的直观性和解释性。农业知识包含如物种、生态系统等层级概念，具有自然的层次结构[21]，如图 2 所示，因此构建的知识图谱具备良好的分层特征。

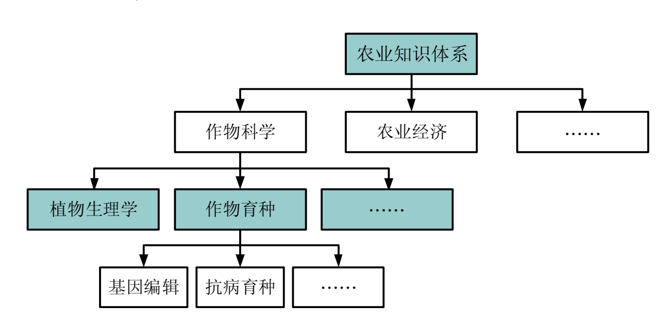


图 2 部分农业知识体系结构

Fig.2 Part of the agricultural knowledge structure

为验证农业知识图谱的层次性，本文选取层聚类系数和模块度两个指标。在图论中，节点的度是指与该节点直接相连的边的数量。节点 i 的度分为入度（In-degree）和出度（Out-degree）。入度表示指向节点 i 边的数量，出度表示从节点 i 指出边的数量。节点聚类系数 $C(i)$ 计算节点 i 的邻居节点之间存在的边与可能存在的边的比例，计算公式如（1）所示。其中，$E(i)$ 是节点 i 的邻居节点之间的边数，k 是节点 i 的度数。

$$C(i)=\frac{2E(i)}{k(i)\times(k(i)-1)} \tag{1}$$

社区模块度用来描述节点间聚合的倾向，较高的模块度表明图中存在明显的社区结构。模块度的定义为公式（2）所示。其中 c 表示社区；m 表示图内总边数；k_c 表示社区内节点度之和；γ 表示分辨率参数，γ 越大代表社区数量越多。

$$Q=\frac{1}{2m}\sum_{c}[e_c-\gamma\frac{k_c^2}{2m}] \tag{2}$$

对已构建的农业知识图谱层次性进行分析，并选择相同节点和边数量的随机图进行对比，其中统计聚类系数的平均值与社区间模块度如表 1 所示。

表 1 农业知识图谱层次性评估

Table 1 Agricultural knowledge graph hierarchy evaluation

编号	图名称	聚类系数	平均模块度
1	农业知识图谱	0.499	0.357
2	随机图-1	0.297	0.014
3	随机图-2	0.307	0.040
4	随机图-3	0.300	0.027

从数据上看，已有农业知识图谱的聚类系数为 0.499，明显高于随机图-1(0.297)、随机图-2(0.307)和随机图-3(0.300)。这表明农业知识图谱中节点更倾向于形成紧密的群组，而随机对照组中的节点则较少形成；已有农业知识图谱的平均模块度为 0.357，远高于随机图-1(0.014)、随机图-2(0.040)和随机图-3(0.027)。这说明农业知识图谱中的社区结构更加显著，而随机对照组中的社区结构较弱，可见农业知识图谱具有更明显的分层结构特征。此外，不同数据集下聚类系数与节点占比关系如图 3 所示。显然，随机图谱中节点主要分布在低聚类区间，相反农业知识图谱较随机图谱相比各聚类系数间分布较均匀，聚类效果显著。因此可认为其产生的分层效果是整体特性，而非是由某些特殊节点呈现出的特征。

4 模型介绍

本部分将结合网络层次表示学习范式（HARP）详细介绍改进算法的具体内容和流程。在分别介绍分层

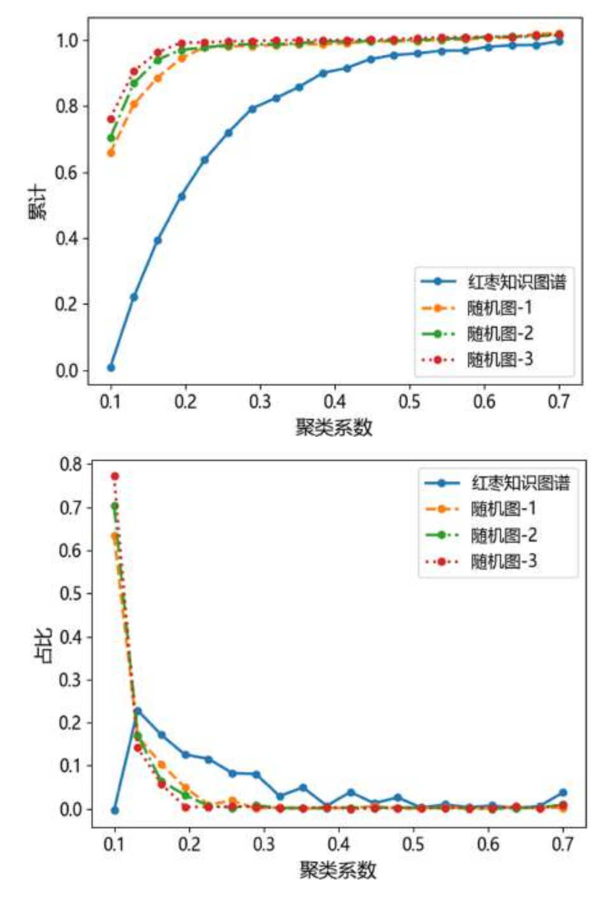


图 3　不同聚类系数下各数据集中节点占比

Fig.3 Node proportion of each dataset under different clustering coefficients

表示算法、改进随机游走算法和 LEIDEN 社区发现算法后，本章将说明如何在 HARP 框架下利用知识图谱中不同维度的结构关系进行快速学习，并尽可能保持原图语义结构和关系的准确性和完整性。

对后文所需符号和定义做出说明：将图表示为 $G=(V, E)$，其中 V 代表点集，E 表示边集；i，j 等表示图中不同顶点，e_{ij} 为连接点 i，j 的边；将图的邻接矩阵记作 $\boldsymbol{A}$，其维数为$|V|$。若 i、j 之间存在边，则A_{ij} 为 e_{ij} 的权重 ω_{ij}，否则 $A_{ij}=0$。

4.1　网络层次表示学习模型

层次表示学习含 3 个步骤[22]：① 图粗粒度化(Graph Coarsening)。将原图 G 通过算法粗粒度化，得到一系列规模逐渐减小的图 $G_0, G_1, \cdots, G_L$。② 表示学习。在规模最小的 G_L上嵌入。由于 $|V|$ 和 $|E|$ 都很小，因此可以快速学习到较好的表示结果。③ 表示提升(Representation Refinement)。通过较小网络的嵌入结果迭代求得下一级网络的嵌入表示。具体而言，对于每个 G_i，将 G_i+1 的嵌入向量作为 G_i 中节点的初始向量，然后继续使用嵌入函数表征 G_i 中每个节点，直到得到 G_0，即原网络的嵌入向量。

4.2　改进随机游走算法

随机游走是从某一起点出发，按照一定概率向邻近节点转移直至到达终点的过程。简单随机游走利用古典概型，定义转移关系如公式（3）。

$$\boldsymbol{B}_{ij}=\begin{cases}\dfrac{\omega_{ij}}{\sum\limits_{j}\omega_{ij}} & A_{ij}\neq 0\ and\ j\neq e\\ 0 & A_{ij}=0\ or\ j=e\end{cases} \tag{3}$$

由于简单随机游走的采样模型利用对称结构定义节点间转移概率，因此无法表示图谱中非对称关系，同时会产生“沉没效应”，使部分重要节点的嵌入结果产生偏移[6]。以图 4 为例，假设“植物”是目标，简单随机游走最快找到“利他素”节点，但显然“陆生植物”和“水生植物”与“植物”更接近。这是由于“利他素”无其他邻居而终止游走，使命中概率变高。可见简单随机游走通常不能反映现实情况和实际需求。

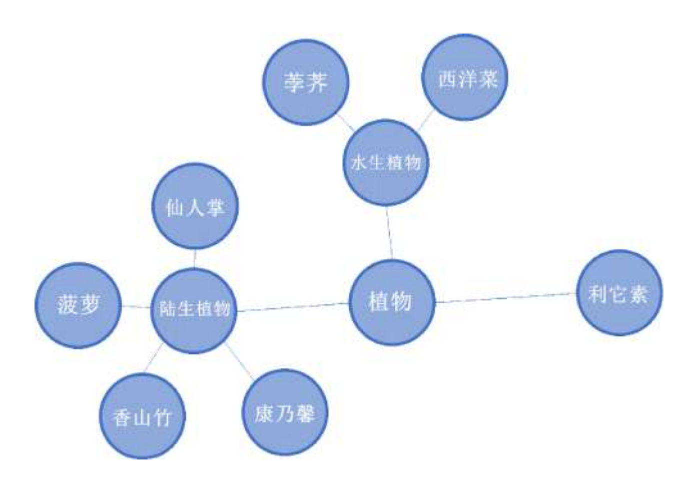


图 4　部分植作物图谱示意

Fig.4 Part of the planting crop graph

为改进上述问题，改进算法以简单随机游走为基础，对采样模型转移矩阵 $\boldsymbol{B}_{ij}$ 做出改进。

DOI：10.13998/j.cnki.issn1002-1248.23-0300

$$\boldsymbol{B}_{ij}=\begin{cases}\dfrac{\omega^2}{D_iD_j} & A_{ij}\neq 0\ and\ i,j\neq e\\ 0 & A_{ij}=0\ or\ i,j=e\end{cases}\qquad(4)$$

其中，$D_i=\sum\omega_{ik}$。定义 i 向 j 转移的概率为 ω_{ij}/D_i，j 接收 i 的概率为 ω_{ji}/D_j。该采样方式可以保证度更高的出发点更容易被接收。不同于简单随机游走，具有阻塞机制的改进随机游走算法的起始点有一定概率被终点拒绝留在原地。

4.3 LEIDEN 聚类算法

LEIDEN 是以 LOUVAIN 算法为基础，基于多层次模块度（Modularity）优化的非重叠社区聚类算法。一个好的划分表现为社区内部节点的相似度较高，而在外部的相似度低。LEIDEN 初始化每个节点为单独社区，尝试将节点 i 分配到邻社区，计算模块度增益。选增益最大的节点加入相邻社区并细化且分区不改变。细化后按节点内权重和更新环权重，区间权重更新为新节点权重并继续迭代直至无改进[23]，示意图如图 5 所示。该算法的优势在于：①可以根据需要定义社区密度；②适用于有向有权图；③不同社区是连通的；④算法通过迭代得到社区划分树，可使用粒度、模块度等中间变量。

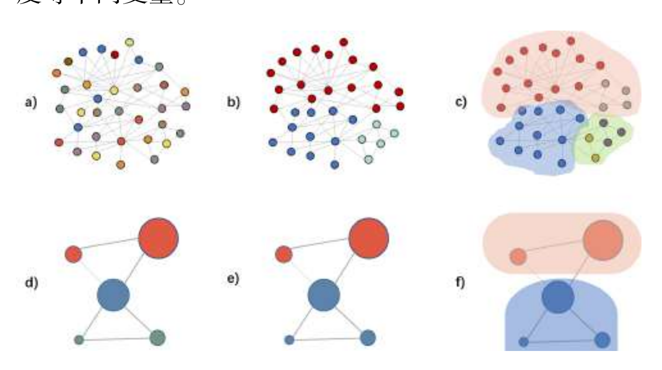


图 5 LEIDEN 聚类算法示意

Fig.5 LEIDEN clustering algorithm

4.4 基于改进随机游走的分层嵌入

为有效学习知识图谱中存在的潜在知识，对于任意图 G，定义目标函数 $f(G)$: $V\rightarrow R^{|V|\times d}$，$d<<|V|$。考虑到采样过程中高维信息缺失及非对称表示等问题，模型通过层次表示学习的方法，利用 LEIDEN 聚类算法和改进随机游走算法对知识图谱的嵌入过程进行优化。其完整的表述为：对于大型知识图谱 $G(V,E)$及其聚类子图 $G_0, G_1,\cdots,G_L$，利用改进随机游走函数 f 迭代得到一系列采样 $\emptyset_{Gi}=f(G_i)$，并最终得到 G 图嵌入表示的过程。

4.4.1 改进随机游走嵌入算法

改进模型针对知识图谱的多关系特征，采用基于关系路径的有偏采样策略，使结果中目标节点与当前节点在语义上更加相近并能有效反映图谱中的非对称关系。对于任意关系序列 $P=T_0\xrightarrow{r_1}\cdots\xrightarrow{r_l}T_l$,$T_i$=range($r_i$)=$domain(r_{i+1})$，定义其沿关系路径 r 游走的概率为：

$$P(r_{i+1}=r)=\frac{ln(\omega_{|r|}+1)}{\sum_{R_v}ln(\omega+1)}\quad v\in T_i\qquad(5)$$

其中，R_v 表示与节点 v 相关的关系集，ω 代表关系权重。上述处理方式可以确保不会忽视偶发关系。对于关系 r，其后续节点选择的概率可以表示为：

$$P(t_i=x\,|\,(h=t_{i-1},r=r_i))=\frac{\omega^2_{(t_{i-1},r_i,x)}}{\sum_h\omega_{(\cdot,r_i,x)}\sum_t\omega_{(t_{i-1},r_i,\cdot)}}\qquad(6)$$

4.4.2 融合 LEIDEN 的分层嵌入模型

为有效保留全局结构的一、二阶相似度，分层嵌入模型采用结合边融合（Edge Collapsing）与星形融合（Star Collapsing）的混合粗化算法对大型图进行分层。边融合是将连接在同一节点的任意两边合并为同一个节点。实验结果表示，其合并顺序不会对结果产生明显的影响[22]。星形融合基于中心节点（Hub）对边缘压缩。由于中心节点周围的点具有相似的结构特征，因此采用星形融合可以很好地保留原图的二阶相似性。在实践中，模型首先对图进行星形融合，然后进行边融合，直至得到一个足够小的图，具体过程如图 6 所示。

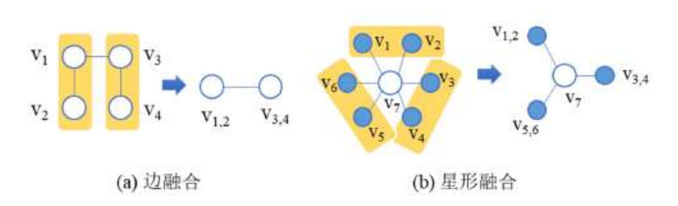


图 6 图粗化算法示意

Fig.6 Graph coarsening algorithm

本文使用 LEIDEN 算法中间变量社区树（Dendrogram）实现图采样。对于各级划分而言，将社区内度

最高的点视为中心节点，逐步合并至节点数小于阈值。

4.4.3　分层随机游走嵌入模型

由于采样方法具有非对称的阻塞机制，对于任意节点 v，其转移的概率和 $\sum_{R_v} P(v,r,x)<1$，阻塞随机游走将其处理成转移概率为 $1-\sum P(v,x)$ 的自环。本模型除顶层节点外，其余节点将以该概率向上层节点转移，顶层节点则为自环，如图 7 所示。

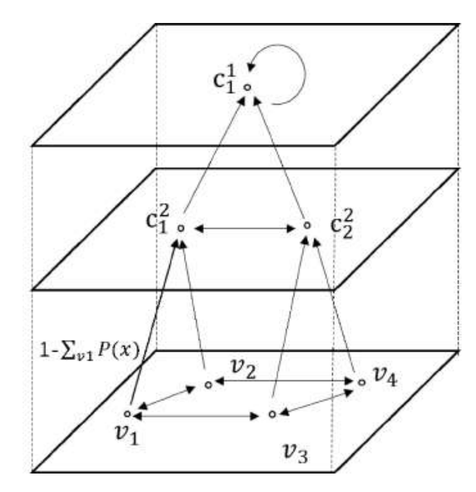


图 7　层间游走采样模型示意

Fig.7 Schematic diagram of interlayer sampling model

由于上层节点是真实存在，因此若某一点与上层节点间存在关系则以该关系类型作为路径嵌入；如果不存在则以统一的虚拟关系嵌入。其目的在最终结果中体现二者的分层关系。

在得到基于关系的有偏抽样路径后，利用 Skip-Gram 学习实体向量，并得到知识图谱的初步嵌入结果。对于三元组 $tri(h,r,t)$，其头部实体 h 的嵌入结果在三元组 emb_h 表示为 h 的初始值与关系嵌入结果 r_h 之和，即 $emb_h=h+r_h$。类似地，尾部实体嵌入表示为 $emb_t=h+rt$。这种表示方法使得嵌入结果更加灵活，并能针对模型特点和需要直接集成到现有的表示学习算法中。模型表示如算法 1（图 8），具体流程如图 9 所示。

算法 1：HRWP 改进模型

输入：知识图谱 G

输出：嵌入结果 $Embeddings$

$P \leftarrow$ LEIDEN(G)

while $|P| < G_V$

$G_{org} \leftarrow G$

$P_{refined} \leftarrow$ DENDROGRAM (G, P)

$G \leftarrow$ AGGREGATE$(G, P_{refined})$

$P \leftarrow \{\{v \mid v \subseteq C,\ v \in G\}, C \in P\}$

while $|G_L| > \varepsilon * G$

$G_L \leftarrow$ EDGE_COL (STAR_COL (G_{org}))

$Corpus$ = FRW_REL (FRW($G_0, G_1, \cdots, G_L$))

$Embeddings$ = SKIP_GRAM($corpus$)

图 8　HRWP 改进模型算法

Fig.8 HRWP improved model algorithm

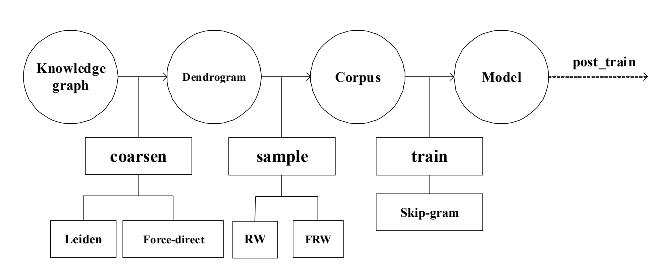


图 9　HRWP 模型总体示意

Fig.9 HRWP model general schematic

5　实验结果分析

为验证本模型的嵌入结果在准确性和速度上的效果，本文选择知识图谱中链接预测任务进行实验。模型代码由 python 语言编写，基于 Pytorch 框架实现。实验主机使用 Intel(R) Core(TM) i7-11800H 处理器及 NVIDIA GeForce RTX 3070。实验中使用 FB15K-237[24] 和 WN18RR[25]数据集进行评估，具体统计信息如表 2 所示。实验将在精度和时间角度与传统模型进行比较。

表 2　实验数据集信息

Table 2 Experimental dataset information

数据集	FB15K-237	WN18RR
实体数量/个	14 541	40 943
关系数量/个	237	11
训练集大小/组	272 115	86 835
验证集大小/组	17 535	3 034

5.1　聚类结果的比较分析

为直观地展示 HRWP 模型中粗化算法保留的结构特征，本文分别将不同训练数据集各层间表示结果通过二维可视化形式表示，如图 10 所示。结果显示训练集 FB15K-237 与 WN18RR 相比具有更强的中心聚集

DOI：10.13998/j.cnki.issn1002-1248.23-0300

性；LEIDEN 聚类算法能更好地保留空间结构，并很好地作为后续层上布局的扩展，而力引导算法（Force-direct，原 HARP 采用的算法）在边融合和星形融合上有更好的效果，但在不同层间结构的保留效果上并不理想。

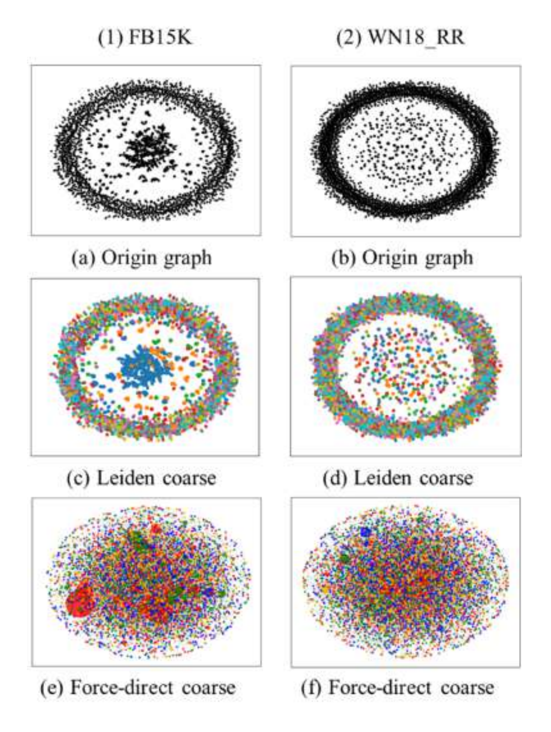


图 10 模型聚类结果可视化

Fig.10 Visualization of model clustering results

图 11 展示了模型中混合粗化方法在测试图中的效果。结果显示，对于不同图第一步粗化操作都能融合 50%。随着粗化过程的继续所有图的规模都以指数规模下降。在第 5 至 8 级后，图中的节点和边的规模均低于 1%。另外，图中可以看出不同测试图中边的融合速度均快于结点的融合速度，但二者收敛趋势相同。这也符合模型边融合与星型融合相结合的特征。

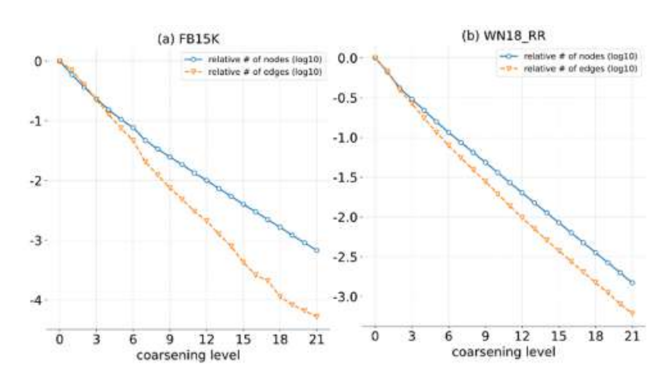


图 11 粗化图的节点 / 边与原始图的比值（取对数）

Fig.11 Ratio of nodes/edges of coarsened graph to the original graph (logarithm)

图 12 为 LEIDEN 算法下不同层级间模块度的变化曲线。从图中可以看出 LEIDEN 作为贪心算法，可以快速收敛至模块度最大值，缺点是结果不具备稳定性。

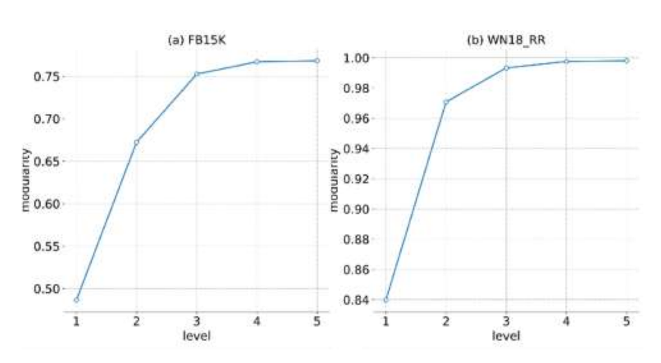


图 12 不同层级间模块度变化

Fig.12 Modularity changes between different levels

5.2 表示时间的比较分析

本节将讨论模型针对不同数据集中对训练时间的改进效果。图 13 中所示的是 HRWP 与对照模型均达到收敛所需时间的示意图。从图上可以看出仅采用 HRWP 所需时间在两数据集上所需时间均远小于对照模型。另外，采用 HRWP 框架的融合模型训练时间基本小于二者训练时间的总和，因此认为 HRWP 模型的结构可用于其他模型的预训练结果，并有效地将缩短其训练时

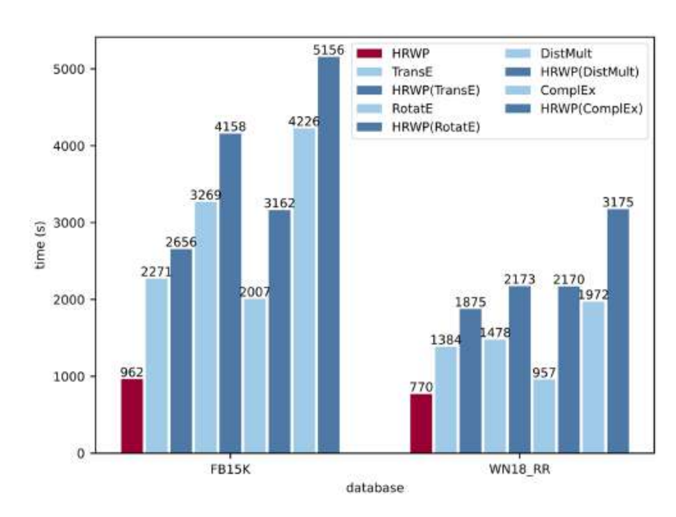


图 13 不同数据集训练时间统计

Fig.13 Training time statistics of different datasets

间。但对于复杂模型（如 DistMult 和 ComplEx），HRWP 训练的结果会降低收敛速度。因此可认为对于多维表示的复杂模型，将表示结果统一对原模型的速度并无提升，因此不适合与 HRWP 模型搭配使用。

为讨论不同采样算法对训练速度的影响，本文选择在农业图谱上选择不同数量的结点进行对比实验。图谱中节点数为 100 到 100 000，并控制节点平均度为 10。从图 14 中可以观察到 HWRP 的运行时间趋势与 FRW 算法相同。与对照方法相比，HWRP 模型中图粗化与延拓过程所带来的时间开销可以忽略不计，特别是当节点规模较大时。此外，根据图中不同组模型采样结果走势可知，HRWP 模型对原算法时间复杂度几乎没有影响。

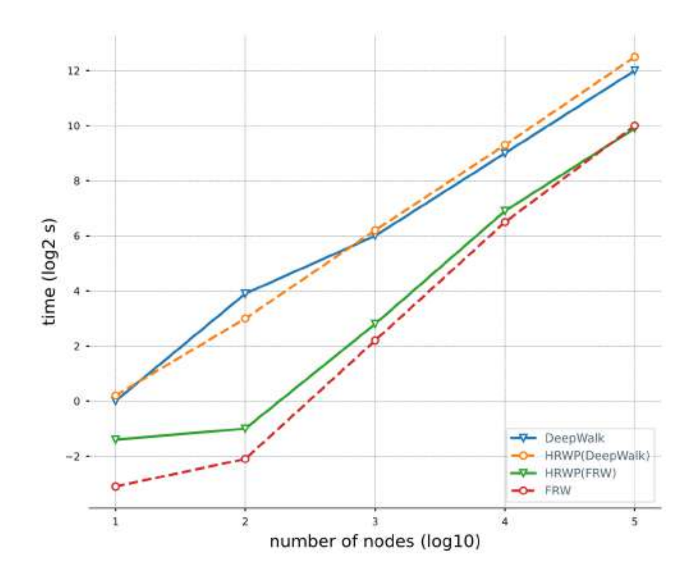


图 14　模型训练时间随节点数量变化分析

Fig.14　Analysis of model training time varying with the number of nodes

5.3　表示效果的比较分析

本节主要比较 HRWP 在 FB15k-237 和 WN18RR 数据集中与其他算法在表示效果上的分析。本文分别使用 MRR（Mean Reciprocal Rank，MRR）和 Hits@*k* [26] 作为判断模型效果的度量。其中 MRR 定义为测试集上三元组倒数的算术平均值，其数值越大模型性能越好。Hits@k代表单个排序位于前 *k* 的三元组比率，其数值越大代表模型效果越好，其中 *k* 通常取 1、3 和 10。实验结果如表 3 所示，对于 FB15k-237 数据集，引入 HRWP 模型后，各指标基本优于或与原始模型持平。结合 HRWP 的 TransE、DistMult、ComplEx 和 RotatE 算法，各指标平均增长了 2%。与神经网络方法比较，快速模型的 MRR 略低于 CompGCN。在 WN18RR 数据集上，非神经网络模型的各指标在加入快速框架后有显著提升。这表明，本文提出的关系路径游走方法使传统算法具备基于神经网络模型的效果，并有效降低训练成本。此外，从表 3 还能看出，引入 HRWP 模型后，不具备识别非对称关系的 TransE 模型进一步提升至与 RotatE 相近。因此可以初步认定模型能够处理非对称关系。

为进一步验证其在复杂关系下的表现性能，本文设计了额外实验，结果如表 4。实验将关系类型进一步划分为 1 对 1（1-1）、1 对多（1-*N*）以及多对多

表 3　FB15K-237 及 WN18RR 数据集上关系预测性能指标对比

Table 3　Comparison of relationship prediction performance indexes in FB15K-237 and WN18RR datasets

项目	FB15K-237				WN18RR			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
TransE	0.292	0.192	0.325	0.478	0.227	0.162	0.233	0.501
TransE-HRWP	0.322	0.232	0.334	0.464	0.252	0.192	0.284	0.510
DistMult	0.241	0.155	0.263	0.419	0.430	0.390	0.440	0.490
DistMult-HRWP	0.284	0.194	0.316	0.463	0.467	0.414	0.461	0.509
RotatE	0.301	0.211	0.331	0.483	0.465	0.428	0.492	0.571
RotatE-HRWP	0.334	0.245	0.349	0.543	0.484	0.438	0.499	0.583
ComplEx	0.248	0.149	0.283	0.423	0.440	0.412	0.463	0.517
ComplEx-HRWP	0.260	0.201	0.281	0.439	0.460	0.409	0.491	0.579
CompGCN	0.343	0.257	0.367	0.523	0.479	0.443	0.494	0.546

DOI：10.13998/j.cnki.issn1002-1248.23-0300

(*N*-*N*)，得到不同算法在 FB15K-237 数据集上关系预测的准确率。结果表明，无论在哪种类型的非对称关系上，融合 HRWP 的 TransE 模型基本表现出了与 RotatE 相近的性能，这进一步证明了模型在处理非对称关系上的优势。

表 4 FB15K-237 上关系预测性能指标对比（Hits@10）

Table 4 Comparison of performance indexes of relationship prediction on FB15K-237(Hits@10)

项目	1-1	1-*N*	*N*-*N*
TransE	0.616	0.463	0.224
TransE-HRWP	0.663	0.417	0.276
RotatE	0.729	0.612	0.302

为验证模型在农业知识图谱上的表现效果，本文选取关系预测任务获得实验数据如表 5 所示。HRWP 模型虽然指标偏低，但 Hits@10 指标与其他模型差距较小，且作为预表示时模型准确性有显著提升。结果表明在准确率要求不高时 HRWP 模型基本满足需求，而对于要求较高的任务其可以与其他模型结合进一步提升表示效果。

表 5 农业知识图谱上性能指标对比

Table 5 Performance index comparison in agricultural knowledge graph

项目	MRR	Hits@1	Hits@3	Hits@10
HRWP	0.201	0.113	0.207	0.454
HRWP-RotatE	0.328	0.232	0.321	0.482
RotatE	0.324	0.228	0.319	0.478
CompGCN	0.365	0.303	0.424	0.526

为验证 HRWP 模型中不同机制对表征学习效果的改善效果，本文采用交叉实验进行要验证。实验分别从框架中去除随机游走模块和层间嵌入单元改为随机嵌入和单层表示学习，并观察其在嵌入时间和准确性地表现，实验结果如表 6 所示。其中 HRWP-H 表示分层随机初始化嵌入，HRWP-RW 表示保留关系路径下随机游走的单层嵌入；RotatE-HRWP 表示完整融合模型。实验中模型嵌入时间如图 15 所示。从结果可以看出，HRWP-H 的表示结果与原始模型的表示结果基本相同，而 HRWP-RW 比原始模型在性能上提高了约 3%。该结果表明引入改进随机游走算法可以有效地表达知识图谱中关系信息。

表 6 交叉模型在 FB15K-237 数据集上的准确性

Table 6 Accuracy of cross model on FB15K-237 dataset

项目	MRR	Hits@1	Hits@3	Hits@10
RotatE	0.301	0.211	0.331	0.483
RotatE-HRWP-H	0.295	0.218	0.329	0.503
RotatE-HRWP-RW	0.339	0.241	0.331	0.503
RotatE-HRWP	0.334	0.245	0.349	0.543

在相同收敛效果下 RotatE 和 RotatE-HRWP-RW 的训练时间均超过 60 分钟，而 RotatE-HRWP-H 和 RotatE-HRWP 的训练时间照对照组均有所下降。这表明基于 HARP 的分层表示学习框架可以有效提升模型的训练速度。交叉实验表明，HRWP 框架中分层嵌入加快了知识图谱表示学习的训练速度，而改进随机游走算法的引入则有效地提高了表示学习的表示性能。

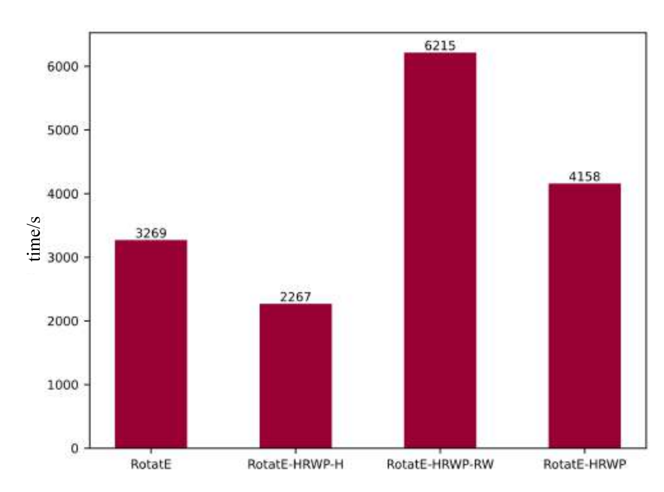


图 15 交叉模型在 FB15K-237 数据集上训练时间

Fig.15 Training time of the cross model on FB15K-237 dataset

6 结　语

本文提出基于 HARP 模型的农业知识图谱分层嵌入模型通过关系路径下的随机游走策略和层间游走相结合，能够有效表示农业知识图谱中关系嵌入的语法和层级关系，同时有效降低训练所需资源的消耗。本文在真实农业知识图谱上对模型嵌入速度和表示效果等方面进行实验，实验数据表明 HRWP 可以有效提高

原始模型性能，同时显著减少表示学习的训练时间。此外，模型具有良好的可扩展性，可以在直接应用于选择其他评分函数的模型。未来研究将更关注对关系对象层级性的讨论和层次分解效果的分析。

参考文献：

[1] 孙坦, 丁培, 黄永文, 等. 文本挖掘技术在农业知识服务中的应用述评[J]. 农业图书情报学报, 2021, 33(1): 4-16.
SUN T, DING P, HUANG Y W, et al. Review on the application and development strategies of text mining in agriculture knowledge services[J]. Journal of library and information science in agriculture, 2021, 33(1): 4-16.

[2] 曹树金, 吴育冰, 韦景竹, 等. 知识图谱研究的脉络、流派与趋势——基于 SSCI 与 CSSCI 期刊论文的计量与可视化[J]. 中国图书馆学报, 2015, 41(5): 16-34.
CAO S J, WU Y B, WEI J Z, et al. History, schools and trend in knowledge map: Investigation and visualization based on SSCI and CSSCI[J]. Journal of library science in China, 2015, 41(5): 16-34.

[3] 徐有为, 张宏军, 程恺, 等. 知识图谱嵌入研究综述[J]. 计算机工程与应用, 2022, 58(9): 30-50.
XU Y W, ZHANG H J, CHENG K, et al. Comprehensive survey on knowledge graph embedding[J]. Computer engineering and applications, 2022, 58(9): 30-50.

[4] 徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述 [J]. 电子科技大学学报, 2016, 45(4): 589-606.
XU Z L, SHENG Y P, HE L R, et al. Review on knowledge graph techniques[J]. Journal of university of electronic science and technology of China, 2016, 45(4): 589-606.

[5] ZHOU C, LIU Y Q, LIU X F, et al. Scalable graph embedding for asymmetric proximity[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 31(1): 456-545.

[6] LI E Z, LE Z Y. Frustrated random walks: A faster algorithm to evaluate node distances on connected and undirected graphs [J]. Physical review E, 2020, 102: 052135.

[7] QIAO L, JIANG L, HAN M, et al. Hierarchical random walk inference in knowledge graphs[C]// The 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. Pisa, Italy: Association for Computing Machinery, 2016.

[8] 赵卓翔, 王轶彤, 田家堂, 等. 社会网络中基于标签传播的社区发现新算法[J]. 计算机研究与发展, 2011, 48(S3): 8-15.
ZHAO Z X, WANG Y T, TIAN J T, et al. A novel algorithm for community discovery in social networks based on label propagation[J]. Journal of computer research and development, 2011, 48(S3): 8-15.

[9] 孙光福, 吴乐, 刘淇, 等. 基于时序行为的协同过滤推荐算法[J]. 软件学报, 2013, 24(11): 2721-2733.
SUN G F, WU L, LIU Q, et al. Recommendations based on collaborative filtering by exploiting sequential behaviors[J]. Journal of software, 2013, 24(11): 2721-2733.

[10] BORDES A, USUNIER N, GARCIA-DURáN A, et al. Translating embeddings for modeling multi-relational data[C]// Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. New York: ACM, 2013: 2787-2795.

[11] FENG J, ZHOU M T, HAO Y, et al. Knowlege graph embedding by flexible translation[J]. arXiv: 1505.05253, 2015.

[12] XIAO H, HUANG M L, HAO Y, et al. TransA: An adaptive approach for knowledge graph embedding[J]. arXiv: 1509.05490, 2015.

[13] SUN Z Q, DENG Z H, NIE J Y, et al. RotatE: Knowledge graph embedding by relational rotation in complex space[J]. arXiv: 1902.10197, 2019.

[14] NICKEL M, TRESP V, KRIEGEL H P. A three-way model for collective learning on multi-relational data[C]// Proceedings of the 28th International Conference on International Conference on Machine Learning. New York: ACM, 2011: 809-816.

[15] YANG B, YUH W T, HE X D, et al. Embedding entities and relations for learning and inference in knowledge bases[J]. arXiv preprint arXiv:1412.6575, 2014.

[16] SCHLICHTKRULL M, KIPF T N, BLOEM P, et al. Modeling relational data with graph convolutional networks[C]//European semantic web conference. Cham: Springer, 2018: 593-607.

[17] SHANG C, TANG Y, HUANG J, et al. End-to-end structure-aware convolutional networks for knowledge base completion[J]. Proceedings of the AAAI conference on artificial intelligence AAAI conference on artificial intelligence, 2019, 33: 3060-3067.

[18] MITCHELL T, FREDKIN E. Never-ending language learning[C]//

DOI：10.13998/j.cnki.issn1002-1248.23-0300

2014 IEEE International Conference on Big Data (Big Data). Piscataway, New Jersey: IEEE, 2015: 1.

[19] DONG X, GABRILOVICH E, HEITZ G, et al. Knowledge vault: A web-scale approach to probabilistic knowledge fusion[C]// Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM, 2014.

[20] 李小凤, 肖帅, 刘希艳, 等. 我国农业类国家标准分类检索浅析[J]. 中国标准化, 2020(7): 67-71, 75.

LI X F, XIAO S, LIU X Y, et al. A brief analysis of the classification retrieval of Chinese national standards for agriculture[J]. China standardization, 2020(7): 67-71, 75.

[21] CHEN H C, PEROZZI B, HU Y F, et al. HARP: Hierarchical representation learning for networks[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1): 336-341.

[22] TRAAG V A, WALTMAN L, VAN ECK N J. From Louvain to Leiden: Guaranteeing well-connected communities[J]. Scientific reports, 2019, 9: 5233.

[23] TOUTANOVA K, CHEN D, PANTEL P, et al. Representing text for joint embedding of text and knowledge bases[C]. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 1499-1509.

[24] DETTMERS T, MINERVINI P, STENETORP P, et al. Convolutional 2D knowledge graph embeddings[C].Proceedings of the AAAI conference on artificial intelligence, 2018, 32(1): 241-249.

[25] VOORHEES E. The TREC-8 question answering track report[C]. Gaithersburg: Proceedings of TREC-8, 2000.

[26] HERLOCKER J L, KONSTAN J A, TERVEEN L G, et al. Evaluating collaborative filtering recommender systems[J]. ACM transactions on information systems, 2004, 22(1): 5-53.

Representation Model of Agricultural Knowledge Graph Based on the HARP Framework

CHEN Caiming[1], FENG Jianzhong[1*], BAI Linyan[2,3], WANG Jian[1], XIE Nengfu[1], ZOU Jun[1]

(1. Agricultural Information Institute of CAAS, Beijing 100081; 2. Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094; 3. International Research Center of Big Data for Sustainable Development Goals, Beijing 100094)

Abstract: [Purpose/Significance] In the era of big data, the volume of data is growing at an exponential rate. One of the most prominent areas affected by this growth is the field of agriculture. The use of agricultural knowledge graphs, which serve as key infrastructures for managing agricultural knowledge, has expanded significantly. However, as the number of nodes and relationships within these graphs increase, so too does their complexity. This complexity gives rise to new challenges in training and representing such large-scale knowledge graphs. It is therefore of great significance to investigate methods for speeding up the embedding process of agricultural knowledge graphs, while preserving their structural integrity and minimizing resource consumption. This research embarks on a novel exploration to address this issue. It stands out from previous studies by concentrating on a hierarchical representation model for agricultural knowledge graphs. The potential impacts of this research on propelling the advancement of the field and on addressing significant real-world problems are substantial. [Method/Process] To confront this challenge, we propose a hierarchical representation model for agricultural knowledge graphs rooted in the HARP framework. Our model leverages the inherent hierarchical features of the

agricultural knowledge graph. It incorporates an improved random walk strategy based on relational paths to semantically model relationship objects within the agricultural knowledge graph. This innovative approach effectively retains the hierarchy and asymmetrical relationship structure of the nodes in the graph, setting our work apart from previous research. The validity of our proposed model is fortified by a strong foundation of theoretical and empirical evidence. [Results/Conclusions] Our experimental results reveal several key findings. First, the hierarchical random walk with path (HRWP) model using the LEIDEN algorithm can preserve the spatial structure more effectively and converge more quickly to the maximum modularity, in comparison to the HARP framework. Second, the fusion model employing HRWP takes less training time than the total training time of both models combined, without significantly affecting the time complexity of the original algorithm. Third, we observed that when traditional algorithms are integrated with HRWP, there is an average improvement of 2% across various indicators, with a substantial enhancement in non-neural network models. Therefore, our proposed model not only accurately represents the agricultural knowledge graph but also effectively reduces the training time. Despite the promising outcomes of our study, there remain areas of potential improvement. One such area is the need for a more detailed discussion on the hierarchical nature of relationship objects in future research. This provides potential avenues for future exploration in this field. The findings of this research carry profound implications for the development of agricultural knowledge management systems, offering an effective approach to handle the burgeoning complexity of knowledge graphs.

Keywords: knowledge graph; random walk; representation learning; the hierarchical random walk with path (HRWP) framework